# Financial Health Assessment for Households in Kenya

1st Samson Kinyanjui
*dept. Computer Science.*
*Dedan Kimathi University of Technology.*
Nyeri, Kenya
sammainah98@gmail.com

2nd Felix Lopuran
*dept. Computer Science)*
*Dedan Kimathi University of Technology.*
Nyeri, Kenya
felixlopuran@gmail.com

3rd Peter Kimanga
*dept. Computer Science*
*Dedan Kimathi University of Technology.*
Nyeri, Kenya
peterkimanga.m@gmail.com

4th Edna Mugoh
*dept. Computer Science*
*Dedan Kimathi University of Technology.*
Nyeri, Kenya
edna.mugoh6@gmail.com

5th Dennis Kiprotich
*dept. Computer Science)*
*Dedan Kimathi University of Technology.*
Nyeri, Kenya
kiprotich403@gmail.com

6th Patrick Gikunda
*dept. Computer Science*
*Dedan Kimathi University of Technology.*
Nyeri, Kenya
patrick.gikunda@dkut.ac.ke

***Abstract*—The analysis of fiscal position represents significant in the encouragement of economic stability and development specifically to the country like Kenya. This work employs the clustering analysis technique in order to assess the Kenyan households' financial literacy based on a financial dataset. These include transaction details, current incomes, expenditures, savings and investment practices among the clients. This paper makes use of K-means clustering technique to investigate the various clusters of households that possibly exist based on their transactional data and assess the financial characteristics and health profiles of the respective clusters. The data used was gathered from 298 distinct households across 5 counties in Kenya. Using correlation analysis and variance threshold, a feature selection method, which aimed at identifying features depicting the greatest extent of clustering tendency was also created. The clustering resulted into eight groups mainly identifiable by transaction direction, transaction value,transaction family, transaction purpose, and mode of transaction. The findings are useful in forecasting and mitigating potential risks within financial planning and management, and may improve financial wellbeing and sustainability at the household level. The project demonstrates how it is possible to implement machine learning methodologies that provide more valuable information at more efficient assessments of the financial well-being of households in Kenya.**

***Index Terms*—clustering, k-means, financial health, households**

## I. Introduction

Due to the recent financial crises of 2007-2008 and the current COVID-19 [1], financial literacy (FL) and financial health have been areas of focus. The capacity for the effective financing in both, the short and the long term is one of the most important aspects of human and social security [2]. By identifying with the concepts in finance, a good access to the funds, and efficient working at the financial aspects forms the characteristics of financial literacy [3].

As the existing unawareness shows, financial literacy is a big issue which potentially puts many people in a vulnerable position every time the economy turns sour. Till May, 2020, more than 70 countries were having a strategic plan to improve financial literacy on the basis of Organization for Economic Cooperation and Development (OECD) suggestions to ensure economic financial resilience and well-being [3].Research consistently shows that financially literate individuals are better equipped to handle emergency expenses, avoid overindebtedness, and make informed investment decisions.

At a micro level, financial literacy correlates with wiser financial decisions across various domains, including market participation, portfolio diversification, and retirement planning. However, such financial illiteracy can lead to increased poverty levels and hamper progressive changes within the economy as exemplified by researches associating the level of people's financial literacy with the 2007 US financial collapse and the effectiveness of electoral changes. This study selected Kenya because the country has a relatively diverse economy and the citizens' awareness of economic matters is also fairly heterogeneous. Though, the financial inclusion strategies have increased the usage of formal financial services still finds it difficult to grasp and maintain the financial wellness [2].

The purpose of this work is attempt to address these gaps by analysing the use of Machine Learning (ML) in evaluating the solvency of households in Kenya [4]. This research aims at establishing quantitative models for assessing an individual's financial health based on more than one type of financial data including transactions, income, expenditure, savings, and investment records. Utilizing clustering algorithm, the project will aim to produce workable forecasts in regards to the financial situations and risks. The outcome of this assessment carries the propensity of helping persons by introducing and maintaining financial sustainability. Furthermore, as the competencies of ML remain more transparent and effective in the field of financial health assessment, this would facilitate the enhancement of the increased utilization of more appropriate strategic models which would result in an invariably more financially literate populace. This remaining part of the paper is organized into the following sections: Literature review section which summarizes research findings from recent literature on the main determinants of financial literacy and well- being, the methodology section discusses the technique(s) that was followed, and the results and discussion section presents the results received with a comprehensive discussion on them. The conclusion and future work summarizes research findings and identifies future research areas, emphasizing the potential of ML methodologies to enhance our understanding of financial health beyond traditional econometric models.

## II. Literature Review

The following review aims at examining the existing literature regarding the application of ML for estimating Kenyan households' financial status, which includes work done in the period between 2019 and 2024. This, concerns the combination of Artificial Intelligence (AI) in business, and financial health evaluation has cropped up as knowledge domain, particularly in the improvement of organizational performance. This work [4] published in 2021 looks how AI can be employed for evaluation and promote well-being and improve health in workplaces. The study by [5] explains how AI holds a future style of offering more efficient, accurate, relevant, and comprehensively analyses the ways of financial assessments.

The article by [6] from the Center of Economic Research, ETH Zurich looks into a very important area of financial literacy on energy and its influence on consumer trends in energy efficiency of home appliances. Using the model-based clustering analysis, the study aimed at identifying three classified clusters of Swiss population based on the level of literacy – low, medium, and high. However, it is worthy of note that the current study establishes huge gaps based on gender, with females being overrepresented in the low and mid-literacy categories as compared with males and underrepresented in the high literacy category. This gender gap should underline the need to adopt more specific measures to help close the literacy gap and as well facilitate the ability to make proper decisions regarding the consumption of energy. In light of such considerations, there is the consequent improvement of energy related financial literacy, not only among the consumers, but also the promotion of policy measures as well as educational initiatives to improve the energy efficiency.

According to [7], the unique study on the heterogeneity of household finance brings a new method to explore the variations of all kinds of households. Prior studies mostly have investigated one or two variables alone to compare change in situation of the households, while this study aims to stress on regression with more than one variable. By applying structural clustering methods of this paper, the researchers are able to establish the complex relationship between the different forms of assets and debts in defining the clusters of households. This innovative method not only enables the specification of the household finances more accurately but also underlines the important role of the real estate and debt variables as the key drivers of the household financial situations. The results of the current study thus support the future use of advanced clustering techniques in developing a much better understanding of the heterogeneity of the households and how best to address them in research in future.

The use of Machine learning algorithms such as neural networks have been used for assessment of financial health and are discussed in [4]. These algorithms can process large volumes of data, identify patterns, and predict future financial health efficiently. Among the benefits they identify is the ability of AI to bring together a comprehensive system of values that complement liquidity ratios, profitability, and cash flow assessments. This strategy provides a holistic look at a firm's financial health in contrast with traditional methods of financial management.

In a case study focused on Perseroan Terbatas (PT) Pos Indonesia (Persero) from 2018 to 2020, [8] introduce an integrated financial health assessment model aimed at preventing company bankruptcy. This study, although not yet widely cited, contributes valuable insights into the practical application of financial health assessment models in real-world scenarios. They propose a parsimonious model that combines financial ratios, trend analyses, and predictive modeling to assess and monitor the financial health of a company . Because of its simplicity and ability to incorporate several financial measures, the model can be beneficial for firms struggling with financial turmoil. Here, the authors stress the need to use both the liquidity, solvency, and profitability ratios in order to get a proper picture of the state of the financial health of the enterprise. The first two areas in the integrated model can then be applied in the context of PT Pos Indonesia as follows: The authors effectively illustrate the possibility of predicting the potential bankruptcy by analyzing the financial information within the given time period. The existence of the model proves the fact that by identifying certain areas of risk, for example, a decrease in sales or an increase in credit, management can undertake appropriate measures to reduce those risks. Another strength of the study is the ability to apply the model to a real business, thus illustrating its possibility for further use. The authors elaborate on important stages of implementing solutions, such as data gathering, model tuning, and outcome analysis. They underscore the fact that model maintenance involves constant assessment and occasional updates because of changing circumstances. They also discuss the issues related to the implementation of such models, for example, the issues of data, the necessity of their update, and the necessity of the involvement of the stakeholders. They put forward that organizations stand to gain from training programs that would improve the management teams' financial literacy to enable their efficient use of the assessment tools.

## III. Methodology

The dataset Financial Diaries - All transactions by [9] comprises of study conducted over time on 298 different households was used. The dataset offered descriptive financial transaction records of households in Kenya from 2012-2013. It includes the following features: date of the transaction, type of transaction family, such as deposits and withdrawals, purpose or description of the transaction, such as savings contributions or loan repayments, mode of the transaction, such as cash, mobile money transfer, or bank transfer, transaction value in Kenyan shillings, this is the monetary value associated with each transaction, and transaction direction this indicates whether money is entering or leaving the household.

Additionally, the dataset included date range features that provide further temporal insights into the financial behavior of households and accounts. These features are first_trx_date_hh (the first transaction date for the household), last_trx_date_hh (the last transaction date for the household), tot_hh_daysofobs (total number of days the household has been observed), tot_hh_monthsofobs (total number of months the household has been observed), first_trx_date_acc (the first transaction date for the account), last_trx_date_acc (the last transaction date for the account), tot_acc_daysofobs (total number of days the account has been observed), transaction_date (date of the transaction), and transaction_year_month (year and month of the transaction). **Fig 1** illustrates the structure of the methodology steps undertaken.

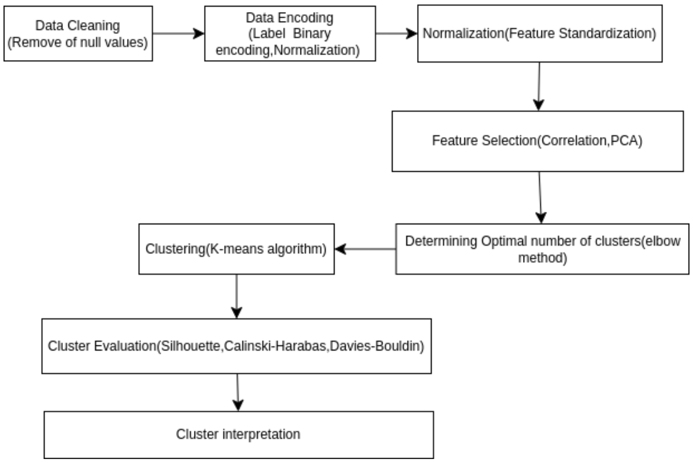
Data Cleaning
(Remove of null values)
Data Encoding
(Label Binary
encoding,Normalization)
Normalization(Feature Standardization)
Feature Selection(Correlation,PCA)
Determining Optimal number of clusters(elbow method)
Clustering(K-means algorithm)
Cluster Evaluation(Silhouette,Calinski-Harabas,Davies-Bouldin)
Cluster interpretation

Fig. 1. Methodology Steps

*a) **Data Pre-processing**:* The initial step in the methodology involved data cleaning, which entailed removing null values to ensure data completeness and accuracy. This was followed by data encoding, where categorical variables were transformed into numeric values to facilitate analysis. Label encoding was applied to trx_class_code and trx_direction, while binary encoding was used for trx_family_code, trx_family_desc, trx_type_desc, and trx_prx_purpose to handle high cardinality and maintain uniqueness $[X = \{x_1, x_2, \ldots, x_n\}$ where $x_i \neq \text{null}$.

*b) **Normalization**:* Normalization was applied to the trx_value_kes feature to scale the numeric values, ensuring that all features contribute equally during the clustering process. This step is crucial to prevent features with larger ranges from dominating the clustering outcome. In order to normalize the variable, we used the formula $X_{\text{norm}} = \frac{X-\mu}{\sigma}$, where $\mu$ is the mean and $\sigma$ is the standard deviation [10].

*c) **Feature Selection**:* Feature selection was performed in multiple stages:

Variance thresholds was the first step, where a threshold of 0.01 was set to remove features with low variance [11]. This process helped in retaining features that are more likely to be useful for clustering. The retained features included trx_class_code, trx_class_desc, various encoded family and type descriptors, trx_mode_code, trx_place_incommunity, trx_direction, trx_value_kes, and trx_value_usd. Features with negligible variance, such as trx_type_desc and trx_prx_purpose, were dropped.

Next, correlation analysis was conducted to identify and remove highly correlated features to avoid redundancy and multicollinearity. This step involved constructing a correlation matrix and dropping one feature from each pair of highly correlated features (correlation is greater than 0.9). Consequently, features like trx_class_desc and several encoded family and purpose descriptors were removed. The correlation coefficient $\rho_{ij}$ between two features $X_i$ and $X_j$ was given calculated using

$$\rho_{ij} = \frac{\sigma_{X_i}\sigma_{X_j}}{\text{Cov}(X_i, X_j)}$$

where $\sigma_{X_i}$ and $\sigma_{X_j}$ are the standard deviations of $X_i$ and $X_j$ respectively, and $\text{Cov}(X_i, X_j)$ is the covariance between $X_i$ and $X_j$ [12] .

Principal Component Analysis (PCA) was then applied to reduce the dimensionality of the data while preserving most of its variance [13]. This was a necessary step since it was critical to reduce the large amount of data and optimize the operation of the clustering function.

*d) **Determining the Optimal Number of Clusters**:* To decide the number of clusters to be created, the method called ‘Elbow Method’ was used. This method involved plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters and identifying the ”elbow point” where the rate of decrease sharply slows. The elbow point in our analysis suggested that four clusters would be optimal [14]. The within-cluster sum of squares (WCSS) for $k$ clusters is given by the formula

$$\text{WCSS}(k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where $C_i$ is the $i$-th cluster and $\mu_i$ is the mean of the $i$-th cluster.

*e) **Clustering**:* The K-means algorithm was then applied to the preprocessed and normalized dataset to perform the

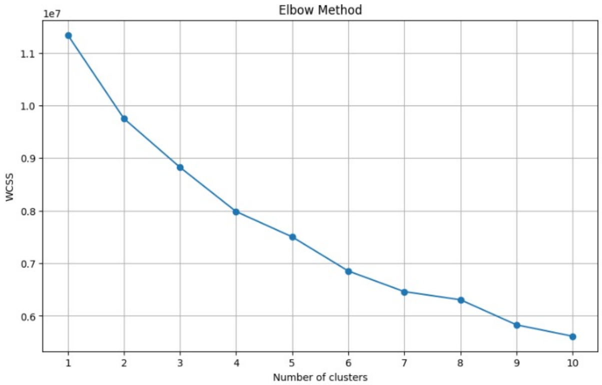

Fig. 2. Elbow Method

clustering [15] . This algorithm partitions the data into K clusters by minimizing the within-cluster variance:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where $\mu_i$ is the centroid of cluster $C_i$.

*f) **Cluster Evaluation:*** The validity of the clustering results was assessed using three metrics:

- The **Davies-Bouldin Index** measured the average similarity ratio of each cluster with the cluster most similar to it. Lower values indicate better-defined clusters [16]. The Davies-Bouldin Index (DBI) is defined as:

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left( \frac{s_i + s_j}{d_{ij}} \right) ]$$

where $s_i$ is the average distance between points in cluster $i$ and $d_{ij}$ is the distance between centroids of clusters $i$ and $j$. For our data, the Davies-Bouldin Index was 1.7979 for four clusters and 1.6265 for eight clusters.

- The **Calinski-Harabasz Index** evaluated the ratio of the sum of between-clusters dispersion and of within-cluster dispersion. Higher values indicate better-defined clusters. Our analysis yielded values of 63,520.74 for four clusters and 51,664.79 for eight clusters.
- The **Silhouette Score** measured how similar a point is to its own cluster compared to other clusters. Scores close to 1 indicate well-clustered data. The Silhouette Score was 0.2664 for four clusters and 0.3497 for eight clusters.

## IV. Results and Discussion

There were thirty one selected features, that showed a correlation that can lead to clusters, these features were selected from a possible 39 features. The selected features were grouped into 8 distinct clusters, this decision was arrived at after trying various values of k, initially k value was taken as k=4, which cluster performance metrics were as shown in Table 1.

TABLE I
Clusters Score when k=4

| Performance Metrics | Value |
|---|---|
| Davies-Bouldin Index(DBI) | 1.78 |
| Calinski-Harabasz Index(CHI) | 63520.74 |
| Silhouette Score(SIL) | 0.27 |

Upon increasing the value of k to 8, the performance of clusters on SIL score increases from 0.27 to 0.35 increases as shown in Table II.

TABLE II
Clusters Score when k=8

| Performance Metrics | Value |
|---|---|
| Davies-Bouldin Index(DBI) | 1.63 |
| Calinski-Harabasz Index(CHI) | 51664.79 |
| Silhouette Score(SIL) | 0.35 |

Increasing the k-value to 8 improved clustering performance, as evidenced by a 0.35 SIL score, 1.63 DBI decrease, and CHI of 51664.79 (Table II). This shows that a higher k-value produced more distinct and well-defined clusters, hence boosting the overall quality and interoperability of clustering.

### *A. Clusters Characteristics*

TABLE III
Dominant Features in all clusters

| Clusters | Features | Feature Number |
|---|---|---|
| All | class, direction, value | 0, 32, 31 |

Table III displays data for eight clusters, each with three significant feature names: ”transaction class description,” ”transaction value(kes),” and ”transaction direction,” representing feature numbers 0, 32, and 31, respectively. These attributes are consistently present in all clusters, indicating that they are important in defining the clusters' characteristics.

Cluster 0, this cluster was characterized by a mix of transaction directions, indicating both inflows and outflows of money. This suggested that households in this cluster were engaged in various financial activities, including both receiving and spending money. The purpose of transactions in this cluster was diverse, including loan repayments, savings contributions, and other financial activities. This pointed to a degree of variability in their financial aspects among households in the given cluster.

Cluster 1, the business people in this cluster seem to spend money in most of the transactions within this cluster suggesting that compared to receiving money, money is likely to flow out of the households. Transactions in this cluster are aimed at a set of objectives, which include

TABLE IV
CLUSTERS FEATURES SUMMARY

| Cluster | Details |
|---|---|
| 0 | Value (KES): 2893.59, Transaction Direction: Inflows and Outflows (avg: 3.0) |
| 1 | Value (KES): 3281.45, Transaction Direction: Inflows and Outflows (avg: 1.23) |
| 2 | Value (KES): 2802.96, Transaction Direction: Inflows and Outflows (avg: 2.63) |
| 3 | Value (KES): 2812.0, Transaction Direction: Inflows (avg: 1.82) |
| 4 | Value (KES): 2796.85, Transaction Direction: Inflows (avg: 3.0) |
| 5 | Value (KES): 3145.55, Transaction Direction: Inflows (avg: 2.24) |
| 6 | Value (KES): 3141.04, Transaction Direction: Inflows and Outflows (avg: 1.35) |
| 7 | Value (KES): 2949.15, Transaction Direction: Inflows (avg: 1.76) |

transfers to savings accounts and meeting the necessary loan repayments, as in Cluster 0. But, the emphasis on withdrawals pointed out a higher level of spending activity compared to the rest of the divisions.

The transaction directions show that this cluster, Cluster 2, experienced inflows as well as outflows. Withdrawals, on the other hand, were larger, which indicate that expenditure was larger than income. The primary means of analysis for this set of transactions was purposes which indicated that the main result expected from these kinds of transactions was the re-payment of loans as well as the contribution to the formation of a savings.

Contrasting Cluster 1 and 2, 0.62 of transactions from Cluster 3 involved money coming in to the household, indicating that more money is coming in than going out. Among the contractual objectives of transactions, savings contributions and loan repayments were common just like in other clusters. This was highly suggestive of a larger income or some amount of financial assistance in this case, given the emphasis on inflows.

Cluster 4, money inflows were the main activity in this cluster, suggesting that households were getting more money than they were spending.Transactions within this cluster included a variety of goals such as savings and loan repayments. With a focus on debt management and savings, the concentration on inflows implied a generally stable financial state. Cluster 5, 6 , and 7 exhibited almost similar characteristics.

All in all, different financial behaviors among households were identified by the clustering analysis. With a balance of inflows and outflows, Cluster 0 demonstrated a variety of financial activity. Predominant outflows from Cluster 1 indicated a spending-focused approach. Even though Cluster 2 displayed both transaction directions, it had higher expenses than revenue, highlighting savings and debt management. Significant inflows were observed in Clusters 3 and 4, indicating either higher income or financial support, and a steady financial position was indicated by the concentration on loan repayments and savings. Clusters 5, 6, and 7, indicates that similar characteristics helped sustain the identified financial trends. Taking all aspects into consideration, all the clusters demonstrated different number of actions related to income and expenditure confirming the presence of a great financial differentiation among households.

## V. CONCLUSION AND FUTURE WORK

In conclusion, this study effectively embodies the use of K-means clustering in evaluating the financial situation of households in Kenya, thus achieving the study's objectives. By identifying distinct clusters founded from the transactional level and using the financial integrity and health status of clusters, it has offered insights into the household financial status for different segments. Thus, it seemed that the concept of clustering, especially the clustering algorithm of K-means, can play a huge role in identifying the distinct patterns of financial behavior and improve the ways of further financial planning and risks management depending on the household group.

Thus, future work should be directed at the following several significant topics to extend the findings of the present research. Firstly, increasing the sample size from the selected households as well as including other more financial variables will help in improving the accuracy. Second, the inclusion of additional exterior data like the social activity on twitter and usage of mobile phones as data input gives a better sample of looking at the financial situation of the households. Last, new personal financial advisory systems derived from the results of machine learning predictions on the base of individual parameters of target households can also enhance their financial behavior. Such advancements are beneficial in helping to make better decisions and apply specific interventions so as to develop a society with the essential capabilities in financial literacy.

## REFERENCES

[1] J. Wullweber, "The covid-19 financial crisis, global financial instabilities and transformations in the financial system," *Global Financial Instabilities and Transformations in the Financial System (September 7, 2020)*, 2020.

[2] N. Ndung'u, "A digital financial services revolution in kenya: The m-pesa case study," *African Economic Research Consortium: Nairobi, Kenya*, pp. 23–44, 2021.

[3] D. B. Santos and H. Gallucci, "Financial illiteracy and customer credit history," *Revista Brasileira de Gestão de Negócios*, vol. 22, no. spe, pp. 421–436, 2020.

[4] T. Krulicky and J. Horak, "Business performance and financial health assessment through artificial intelligence," *Ekonomicko-manazerske spektrum*, vol. 15, no. 2, pp. 38–51, 2021.

[5] W. Heo, J. M. Lee, N. Park, and J. E. Grable, "Using artificial neural network techniques to improve the description and prediction of household financial ratios," *Journal of Behavioral and Experimental Finance*, vol. 25, p. 100273, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214635019302230

[6] N. Kumar, "A model-based clustering approach for analyzing energy-related financial literacy and its determinants," *Available at SSRN 3328468*, 2019.

[7] Y. Hwang, Y. Lee, and F. J. Fabozzi, "Identifying household finance heterogeneity via deep clustering," *Annals of Operations Research*, vol. 325, no. 2, pp. 1255–1289, 2023.

[8] P. Kusuma, F. I. Putra, and T. A. Perdana, "Parsimonious: Initiation of integrated financial health assessment model in preventing company bankruptcy," *Journal of Business and Management Review*, vol. 3, no. 4, pp. 337–358, 2022.

[9] F. Kenya, D. D. Data, and B. F. Associates, "Financial Diaries - All Transactions," 2015. [Online]. Available: https://doi.org/10.7910/DVN/JF8YST

[10] M. Y. Prostov, M. M. Suarez-Alvarez, and Y. I. Prostov, "Properties of the sample estimators used for statistical normalization of feature vectors," *Data mining and knowledge discovery*, vol. 29, no. 6, pp. 1815–1837, 2015.

[11] Z. Hou, Q. Hu, and W. L. Nowinski, "On minimum variance thresholding," *Pattern Recognition Letters*, vol. 27, no. 14, pp. 1732–1743, 2006.

[12] Y. Zu, W. Fan, J. Zhang, Z. Li, and M. Ohsaki, "Investigation of equivalent correlation coefficient based on the mehler's formula," *Engineering Computations*, vol. 36, no. 4, pp. 1169–1200, 2019.

[13] S. Janićijević, V. Mizdraković, and M. Kljajić, "Principal component analysis in financial data science," *Advances in principal component analysis. London: IntechOpen*, pp. 113–138, 2022.

[14] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip Journal on Wireless Communications and Networking*, vol. 2021, pp. 1–16, 2021.

[15] Y. Zhao and X. Zhou, "K-means clustering algorithm and its improvement research," in *Journal of Physics: Conference Series*, vol. 1873, no. 1. IOP Publishing, 2021, p. 012074.

[16] G. Ghufron, B. Surarso, and R. Gernowo, "The implementations of k-medoids clustering for higher education accreditation by evaluation of davies bouldin index clustering," *Jurnal Ilmiah KURSOR*, vol. 10, no. 3, 2020.